UNITED STATES PATENT AND TRADEMARK OFFICE

IFW

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/822,310	04/12/2004	Andre Lavoie	03883-P0022A	2574

7590 09/22/2006

Andre Lavoie
396 Beacon Street
Boston, MA 06902

EXAMINER	
SINGH, RACHNA	

ART UNIT	PAPER NUMBER
2176	

DATE MAILED: 09/22/2006

Please find below and/or attached an Office communication concerning this application or proceeding.

3

Office Action Summary	Application No.	Applicant(s)	
	10/822,310	LAVOIE ET AL.	
	Examiner	Art Unit	
	Rachna Singh	2176	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *12 April 2004*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-29* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-29* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *12 April 2004* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date *10/07/04*.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

1. This action is responsive to communications: Application filed on 04/12/04.

2. Claims 1-29 are pending. Claims 1, 28, and 29 are independent claims.

Priority

3. Applicant's claim for the benefit of a prior-filed application under 35 U.S.C. 119(e) or under 35 U.S.C. 120, 121, or 365(c) is acknowledged. Applicant has not complied with one or more conditions for receiving the benefit of an earlier filing date under 35 U.S.C. 120 as follows:

Regarding Provisional Application 60/461,386, the later-filed application must be an application for a patent for an invention which is also disclosed in the prior application (the parent or original nonprovisional application or provisional application). The disclosure of the invention in the parent application and in the later-filed application must be sufficient to comply with the requirements of the first paragraph of 35 U.S.C. 112. See *Transco Products, Inc. v. Performance Contracting, Inc.*, 38 F.3d 551, 32 USPQ2d 1077 (Fed. Cir. 1994).

The disclosure of the prior-filed provisional application, Application No. 60/461,386, fails to provide adequate support or enablement in the manner provided by the first paragraph of 35 U.S.C. 112 for one or more claims of this application.

Provisional Application 60/461,386 is drawn to a Powered rotary board turner which is not related to a financial document change identifier.

Furthermore, the application Applicant's claim for the benefit of a prior-filed application under 35 U.S.C. 120, 121, or 365(c) is acknowledged. Applicant has not complied with one or more conditions for receiving the benefit of an earlier filing date under 35 U.S.C. 120 as follows: An application for patent for an invention disclosed in the manner provided by the first paragraph of **section 112** of this title in an application previously filed in the United States, or as provided by **section 363** of this title, which is filed by an inventor or inventors named in the previously filed application shall have the same effect, as to such invention, as though filed on the date of the prior application. The inventor(s) named in Provisional Application 60/461,386 do not match any of the names of the inventor(s) of the current application.

Applicant's claims for the benefit of Provisional Application 60/462,065 is acknowledged.

Information Disclosure Statement

4. The information disclosure statement (IDS) submitted on 10/07/04 has been considered by the examiner.

Claim Objections

5. Claims 4-6, 7-8, 13-15, and 22-24 are objected to under 37 CFR 1.75(c), as being of improper dependent form for failing to further limit the subject matter of a previous claim. Applicant is required to cancel the claim(s), or amend the claim(s) to place the claim(s) in proper dependent form, or rewrite the claim(s) in independent form.

Regarding claim 4, claim 1, from which claim 4 depends, requires the generation delta data indicative of ***at least one of change and percentage change***. Claim 1 only requires one of change or percentage change, not both. Thus claim 4 fails to further limit the subject matter of claim 1 because the claim requires the generation of only one of a change or percentage change.

Claims 5-6 are objected to for fully incorporating the deficiencies of their base claim from which they depend.

Regarding claims 7-8, claim 1, from which claims 7-8 depend, requires the generation delta data indicative of ***at least one of change and percentage change***. Claim 1 only requires one of change or percentage change, not both. Thus claims 7-8 fail to further limit the subject matter of claim 1 because the claim requires the generation of only one of a change or percentage change.

Regarding claim 13, claim 12, from which claim 13 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular data***. Thus although claim 13 recites displaying the additions, deletions, and substitutions data as visually distinct from the tabular data, claim 12 only requires that one of these data be displayed; therefore, claim 13 fails to further limit the subject matter of claim 12.

Claim 14 is objected to for fully incorporating the deficiencies of its base claim from which it depends.

Regarding claim 15, claim 11, from which claim 15 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular data***. Thus although claim 15 recites displaying the additions, deletions, and substitutions data in a variety of manners, claim 11 only requires that one of these data be displayed. Therefore, claim 15 fails to further limit the subject matter of claim 11.

Regarding claim 22, claim 21, from which claim 22 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular data***. Thus although claim 22 recites displaying the additions, deletions, and substitutions data as visually distinct from the tabular data, claim 21 only requires that one of these data be displayed. Therefore, claim 22 fails to further limit the subject matter of claim 21.

Claim 23 is objected to for fully incorporating the deficiencies of its base claim from which it depends.

Regarding claim 24, claim 20 from which claim 24 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular data***. Thus although claim 24 recites displaying the additions, deletions, and substitutions data in a variety of manners, claim 20 only requires that one of these data be displayed. Therefore, claim 24 fails to further limit the subject matter of claim 20.

Claim Rejections - 35 USC § 101

6. 35 U.S.C. 101 reads as follows:

> Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title.

7. Claims 1-2 and 9-29 are rejected under 35 U.S.C. 101 because the claimed invention is directed to non-statutory subject matter.

Independent claim 1 is drawn to an apparatus for generating a comparison of related subject matter found in two different financial documents. Although the claim recites "to generate tabular delta data indicative of at least one of change and percentage change between the related subject matter of said first document tabular data and said second document tabular data", the tabular delta data is not necessarily

made available to a user and therefore remains in the abstract. Thus, claim 1 fails to produce a tangible result. In order for a claim to be drawn to statutory subject matter, it must be capable of producing a concrete, useful, and tangible result. In this case, although the result is concrete and useful, it is not tangible. Appropriate correction is required.

Dependent claims 2 and 9-27 do not appear to cure this problem and are therefore rejected under 35 U.S.C. 101 for fully incorporating the deficiencies of their base claim from which they depend.

It is noted that dependent claims 3-6 recite features where the tabular delta data is delivered to a user interface, thus they are drawn to statutory subject matter.

Independent claims 28-29, like claim 1, also recite "to generate tabular delta data indicative of at least one of change and percentage change between the related subject matter of said first document tabular data and said second document tabular data". The claimed tabular delta data is not necessarily made available to a user and therefore remains in the abstract. Thus, claim 1 fails to produce a tangible result. In order for a claim to be drawn to statutory subject matter, it must be capable of producing a concrete, useful, and tangible result. In this case, although the result is concrete and useful, it is not tangible. Appropriate correction is required.

8. Further, to expedite a complete examination of the instant application the claims rejected under 35 U.S.C. 101 (nonstatutory) above are further rejected as set forth below in anticipation of Applicant amending these claims to claim statutory subject matter.

Claim Rejections - 35 USC § 102

9. The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (e) the invention was described in (1) an application for patent, published under section 122(b), by another filed in the United States before the invention by the applicant for patent or (2) a patent granted on an application for patent by another filed in the United States before the invention by the applicant for patent, except that an international application filed under the treaty defined in section 351(a) shall have the effects for purposes of this subsection of an application filed in the United States only if the international application designated the United States and was published under Article 21(2) of such treaty in the English language.

10. Claim 1, 9-11, 16, 18-20, 25, and 28-29 are rejected under 35 U.S.C. 102(e) as being anticipated by Gay, US 6,792,145 B2, 09/14/04 (filed on 06/08/01)

Regarding claim 1, Gay teaches a pattern recognition process for text document interpretation. Gay teaches extracting textual and tabular data from financial documents. A comparison is made between the character strings of the financial document and the character strings provided in the previous financial documents which meets the preamble, ***an apparatus for generating a comparison of related subject matter found in two different financial documents***. See abstract. Gay teaches his

invention is directed to SEC documents such as 10-Q or 10-K financial documents which contain character strings in tabular form. See column 1, lines 35-45 and column 2, lines 23-52. Comparisons are made between a raw SEC document containing tabular information that has been downloaded from a website and a new SEC financial document which also contains tabular information which meets the limitations, ***a first document with at least a portion of said first document in a tabular data format; a second document with at least a portion of said second document in a tabular data format, said second document being a variation of said first document***. See column 3, lines 35-67 and column 4, lines 1-38. Gay teaches receiving the first and second document via a website which meets the limitation, ***a processor for receiving said first document and said second document***. See column 3, lines 35-67 and column 4, lines 1-38. Gay further teaches extracting a first valid character string from a previously existing financial document and comparing each string in a first/old document to the character strings in the new/second financial document which meets the limitation, ***a comparator executing on said processor for comparing said first document tabular data to related subject matter of said second document tabular data***. See figure 1, column 4, lines 14-67 and column 5, lines 1-40. Gay teaches the comparison of the two documents results in the creation of a second matrix of character strings provided on a second plane in the database including those textual strings that are not included in the first matrix of textual strings (from the first document) which meets the limitation ***generate tabular delta data indicative of at least one of change and percentage change between the related subject matter of said first document***

tabular data and said second document tabular data. See columns 5, lines 40-67 and column 6, lines 1-54.

EXAMINER NOTE: Determining which textual strings are new or not included in the first matrix of textual strings representing the first document and forming a second matrix is generating tabular delta data indicative of a "change" because it is identifying a new textual string in the second financial document which is considered a "change".

Regarding claim 28, claim 28 is drawn to a system for the apparatus claimed in claim 1, and therefore is rejected under the same rationale used in claim 1 above.

Regarding claim 29, claim 29 is drawn to a method for the apparatus claimed in claim 1, and therefore is rejected under the same rationale used in claim 1 above.

Regarding claim 9, Gay teaches comparing character strings provided in the previous financial document with the character strings in the second financial document which meets the limitation ***compare sections of the first document tabular data with related subject matter sections of said second document tabular data based on at least one of tables, graphs, columns, rows, time units, idea units and line items.*** See figure 1, column 4, lines 14-67 and column 5, lines 1-40. Examiner Note: Line items are being interpreted as the character strings.

Regarding claim 10, Gay teaches the first and second document tabular data contains text data and the comparator generates the text/tabular delta data. See figure 1, column 2, lines 24-52, column 3, lines 35-66 and column 4.

Regarding claim 11, Gay teaches the delta data can include data that has been added in the new financial document. See column 2, lines 1-15 and column 9,l ines 59-62.

Regarding claim 16, Gay teaches comparing character strings provided in the previous financial document with the character strings in the second financial document which meets the limitation ***compare sections of the first document text/tabular data with related subject matter sections of said second document text/tabular data based on at least one of tables, graphs, columns, rows, time units, idea units and line items.*** See figure 1, column 4, lines 14-67 and column 5, lines 1-40. Examiner Note: Line items are being interpreted as the character strings.

Regarding claim 18, Gay teaches the first and second documents comprise data in a text format. See columns 1-2. Gay further teaches these documents include one or more lines of textual material and one or more columns of data associated with each line of textual material. See column 1, lines 35-46. The textual strings are separated into a separate column from the columns of numerical data. Before comparing the first

document to the second document, a first valid character string is extracted from the old/original document. See column 4, lines 14-38.

Regarding claim 19, Gay further teaches extracting a first valid character string from a previously existing financial document and comparing each string in a first/old document to the character strings in the new/second financial document. See figure 1, column 4, lines 14-67 and column 5, lines 1-40. Gay teaches the comparison of the two documents results in the creation of a second matrix of character strings provided on a second plane in the database including those textual strings that are not included in the first matrix of textual strings (from the first document) which meets the limitation ***generate text delta data***. See columns 5, lines 40-67 and column 6, lines 1-54.

Regarding claim 20, Gay teaches the delta data can include data that has been added in the new financial document. See column 2, lines 1-15 and column 9,l ines 59-62.

Regarding claim 25, Gay teaches comparing character strings provided in the previous financial document with the character strings in the second financial document which meets the limitation ***compare sections of the first document text/tabular data with related subject matter sections of said second document text/tabular data based on at least one of tables, graphs, columns, rows, time units, idea units and***

line items. See figure 1, column 4, lines 14-67 and column 5, lines 1-40. Examiner Note: Line items are being interpreted as the character strings.

Claim Rejections - 35 USC § 103

11. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

12. Claims 2-6, 12-15, 17, 21-24 and 26-27, are rejected under 35 U.S.C. 103(a) as being unpatentable over Gay, US 6,792,145 B2, 09/14/04 (filed on 06/08/01) in view of Horton, US 2004/0230892 A1, 11/18/04 (filed 03/17/04, provisional application filed on 03/17/03).

Regarding claim 2, although Gay teaches storing the tabular data and tabular delta data in separate planes in a database, Gay does not disclose a user interface in communication with a processor for delivering at least one of said tabular data and tabular delta data. However, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously which meets the limitation, ***a user***

interface for delivering at least one of said tabular data and tabular delta data. See page 1, paragraphs [0012]-[0019] and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

Regarding claim 3, Gay does not teach the tabular delta data is delivered on a user interface as visually distinct from the tabular data. However, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously wherein the differences are highlighted in order to make it easy to find the differences which meets the limitation, ***wherein said tabular delta data is delivered on a user interface as visually distinct from the tabular data***. See page 1, paragraphs [0012]-[0019] and figure 1. Highlighted the differences by italicizing certain words is providing a means to visually distinct the delta data from the tabular data.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and

changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

Regarding claim 4, Gay does not teach the tabular delta data is delivered on a user interface as visually distinct from the tabular data in one manner and the percentage change is represented in a second manner. However, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously wherein the differences are highlighted in order to make it easy to find the differences which meets the limitation, ***wherein said tabular delta data is delivered on a user interface as visually distinct from the tabular data in a first manner***. See page 1, paragraphs [0012]-[0019] and figure 1. Highlighted the differences by italicizing certain words is providing a means to visually distinct the delta data from the tabular data. EXAMINER NOTE: Claim 1 requires the generation delta data indicative of ***at least one of change and percentage change***. In this instance, Examiner has relied upon Gay's teachings of generating delta data indicative of a change, not a percentage change since claim 1 only requires one of change or percentage change, not both.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

Regarding claim 5, Gay does not teach displaying a plurality of visually distinct tabular delta data; however, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously wherein the differences are highlighted in order to make it easy to find the differences which meets the limitation, ***a plurality of visually distinct tabular delta data***. page 1, paragraphs [0012]-[0019] and figure 1. Figure 1 displays multiple drafts indicating a plurality of differences.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a

simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

Regarding claim 6, Gay does not teach that the tabular delta data delivered on the user interface is chronicled by at least one of numeric, alphabetic, alphanumeric, and consecutive sequence units. However, Horton teaches delivering tabular delta data chronicled by a draft number relating to the version of the document. See figure 1.

Regarding claim 12, Gay does not teach a user interface for delivering at least one of said additions, deletions, substitutions, and text/tabular data; however, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously wherein any additions, deletions, and substitutions are highlighted which meets the limitation, ***a user interface for delivering at least one of said tabular data and tabular delta data***. See page 1, paragraphs [0012]-[0019], page 3, paragraph [0069], and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a

simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

Regarding claim 13, Gay does not teach the additions, deletions, and substitutions data are visually distinct from the tabular data; however, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously wherein any additions, deletions, and substitutions are highlighted which meets the limitation, ***wherein said additions, deletions, and substitutions data is delivered on said user interface as visually distinct from said text/tabular data***. See page 1, paragraphs [0012]-[0019], page 3, paragraph [0069], and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

EXAMINER NOTE: Claim 12, from which claim 13 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular data***. Thus although claim 13 recites displaying the additions, deletions, and

substitutions data as visually distinct from the tabular data, claim 12 only requires that one of these data be displayed; therefore, Examiner is relying on the fact that Horton displays edited data in a different manner such as by italics.

Regarding claim 14, Gay does not teach the additions, deletions, and substitutions data are displayed in a third, fourth, and fifth manner respectively; however, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously wherein any additions, deletions, and substitutions are highlighted which meets the limitation, ***wherein said visually distinct additions data is represented in a third manner***. See page 1, paragraphs [0012]-[0019], page 3, paragraph [0069], and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

EXAMINER NOTE: Claim 12, from which claim 14 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular***

data. Thus although claim 14 recites displaying the additions, deletions, and substitutions data in a variety of manners, claim 12 only requires that one of these data be displayed, thus Examiner is relying on the fact that Horton displays edited data in a different manner such as by italics.

Regarding claim 15, Gay does not teach that one of the additions, substitutions, or deletions data delivered on the user interface is chronicled by at least one of numeric, alphabetic, alphanumeric, and consecutive sequence units. However, Horton teaches delivering tabular delta data, indicative of changes made to the document, are chronicled by a draft number relating to the version of the document. See figure 1. EXAMINER NOTE: Claim 11, from which claim 15 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular data***. Thus although claim 15 recites displaying the additions, deletions, and substitutions data in a variety of manners, claim 11 only requires that one of these data be displayed.

Regarding claim 17, Gay does not teach integrated at least two of the tabular delta data, text/tabular delta data, tabular data, and text/tabular data for delivery on a user interface. Horton teaches integrating tabular delta data and tabular data for delivery on a user interface as depicted in claim 1. Horton teaches a system and method for document project management in which the original portion of a document

and each of a plurality of proposed revisions are displayed simultaneously. See page 1, paragraphs [0012]-[0019] and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

Regarding claim 21, Gay does not teach a user interface for delivering at least one of said additions, deletions, substitutions, and text/tabular data; however, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously wherein any additions, deletions, and substitutions are highlighted which meets the limitation, ***a user interface for delivering at least one of said tabular data and tabular delta data***. See page 1, paragraphs [0012]-[0019], page 3, paragraph [0069], and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the

differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

Regarding claim 22, Gay does not teach the additions, deletions, and substitutions data are visually distinct from the tabular data; however, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously wherein any additions, deletions, and substitutions are highlighted which meets the limitation, ***wherein said additions, deletions, and substitutions data is delivered on said user interface as visually distinct from said text/tabular data***. See page 1, paragraphs [0012]-[0019], page 3, paragraph [0069], and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

EXAMINER NOTE: Claim 21, from which claim 22 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular data***. Thus although claim 22 recites displaying the additions, deletions, and substitutions data as visually distinct from the tabular data, claim 21 only requires that one of these data be displayed; therefore, Examiner is relying on the fact that Horton displays edited data in a different manner such as by italics.

Regarding claim 23, Gay does not teach the additions, deletions, and substitutions data are displayed in a third, fourth, and fifth manner respectively; however, Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously wherein any additions, deletions, and substitutions are highlighted which meets the limitation, ***wherein said visually distinct additions data is represented in a third manner***. See page 1, paragraphs [0012]-[0019], page 3, paragraph [0069], and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a

simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

EXAMINER NOTE: Claim 21, from which claim 23 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular data***. Thus although claim 23 recites displaying the additions, deletions, and substitutions data in a variety of manners, claim 21 only requires that one of these data be displayed, thus Examiner is relying on the fact that Horton displays edited data in a different manner such as by italics.

Regarding claim 24, Gay does not teach that one of the additions, substitutions, or deletions data delivered on the user interface is chronicled by at least one of numeric, alphabetic, alphanumeric, and consecutive sequence units. However, Horton teaches delivering tabular delta data, indicative of changes made to the document, are chronicled by a draft number relating to the version of the document. See figure 1.

EXAMINER NOTE: Claim 20, from which claim 24 depends, recites delivering ***at lest one of said additions data, deletions data, substitutions data, and text/tabular data***. Thus although claim 24 recites displaying the additions, deletions, and substitutions data in a variety of manners, claim 20 only requires that one of these data be displayed.

Regarding claim 26, Gay does not teach integrated at least two of the tabular delta data, text/tabular delta data, tabular data, and text/tabular data for delivery on a

user interface. Horton teaches integrating tabular delta data and tabular data for delivery on a user interface as depicted in claim 1. Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously. See page 1, paragraphs [0012]-[0019] and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

Regarding claim 27, Gay does not teach integrated at least one of text/tabular delta data and text/tabular data for delivery on a user interface. Horton teaches integrating tabular delta data and tabular data for delivery on a user interface as depicted in claim 1. Horton teaches a system and method for document project management in which the original portion of a document and each of a plurality of proposed revisions are displayed simultaneously. See page 1, paragraphs [0012]-[0019] and figure 1.

It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Horton's display of a portion of the original document and changes to that portion in a graphical user interface in Gay's system for storing the differences between financial documents in a database because it enables a user to simultaneously view the differences between various versions of the same document. This was desirable at the time of the invention in order to provide a user with a simultaneous, side-by-side comparison of the differences between documents. See page 1, paragraphs [0003]-[0015].

13. Claims 7-8 are rejected under 35 U.S.C. 103(a) as being unpatentable over Gay, US 6,792,145 B2, 09/14/04 (filed on 06/08/01) in view of Horton, US 2004/0230892 A1, 11/18/04 (filed 03/17/04, provisional application filed on 03/17/03), as applied to claim 2 above, and further in view of Zilberman, US 2006/0167772 A1, 07/27/06 (filed 10/30/02, provisional application filed on 10/30/02).

Regarding claim 7, neither Gay nor Horton teaches inserting a graphic into the tabular delta data indicative of change magnitude for each change between related subject matter of the first tabular data and the second document tabular data; however, Zilberman teaches an electronic interpretation of financials in which financial inputs related to an entity are evaluated against predetermined values. See abstract, page 1, paragraphs [0006]-[0011]. Zilberman's system includes graphics capabilities so that in addition to outputting text, graphs and charts can be output to illustrate the evaluated

relationships such as the change and percentage change between previous periods which meets the limitation ***inserting a graphic into the tabular delta data indicative of change magnitude for each change between related subject matter of the first tabular data and the second document tabular data.*** See page 6, paragraph [0068]. It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Zilberman's insertion of a graphic depicting the change between financial information in the system of Gay/Horton because it would visually display comparisons of information with previous periods, industry averages, etc. See page 6, paragraph [0068].

Regarding claim 8, neither Gay nor Horton teaches the graphic is comprised of at least one of graphs, charts, statistics, and images. Zilberman's system includes graphics capabilities so that in addition to outputting text, graphs and charts can be output to illustrate the evaluated relationships. See page 6, paragraph [0068]. It would have been obvious to a person of ordinary skill in the art at the time of the invention to incorporate Zilberman's insertion of a graphic depicting the change between financial information in the system of Gay/Horton because it would visually display comparisons of information with previous periods, industry averages, etc. See page 6, paragraph [0068].

Conclusion

14. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

a. Forster US 2004/0220980 A1

b. Guy et al. US 2005/0154664 A1

c. Weinberg et al. US 2002/0116417 A1

d. Brandenberger US 2004/0054967 A1

e. Schulze et al. US 6,446,072 B1

f. Sites US 6,324,555 B1

g. Ferguson et al. US 6,336,094 B1

h. Chow et al. US 6,029,175

i. Cooper, James, et al, Detecting Similar Documents Using Salient Terms, November –9, 2002 CIKM '02. Pages 245-251.

15. Any inquiry concerning this communication or earlier communications from the examiner should be directed to Rachna Singh whose telephone number is 571-272-4099. The examiner can normally be reached on M-F (8:30AM-6:00PM).

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Heather Herndon can be reached on 571-272-4136. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

Rachna Singh
09/11/06

IFW

PATENT
03883-P0022A WWW/TMO

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants	Andre Lavoie, *et al.*
Serial No. 10/822,310	Filing Date: April 12, 2004
Title of Application:	Financial Document Change Identifier
Confirmation No. 2574	Art Unit: 2176
Examiner	

Commissioner for Patents
Post Office Box 1450
Alexandria, VA 22313-1450

Information Disclosure Statement by Applicants

As a means of complying with the duty of disclosure set forth in 37 CFR §1.56, Applicants list the following references which were cited in the parent case, Serial No. 60/461,386. The examiner is requested to inspect that file if he wishes to see the full texts of the cited art.

U.S. Patent Documents				
Exam. Initials	Class/ Subclass.	Document No.	Date	Name
	364/408	5,502,637	3/96	Beaulieu et al.

Mailing Certificate: I hereby certify that this correspondence is today being deposited with the U.S. Postal Service as *First Class Mail* in an envelope addressed to: Commissioner of Patents and Trademarks; Post Office Box 1450; Alexandria, VA 22313-1450.

October 4, 2004

Tamara L. Millikan

The listed patents pertain in a general way to the subject matter of the application, but are not necessarily considered to be analogous prior art.

Respectfully submitted,

September 30, 2004

Wesley W. Whitmyer, Jr., Registration No. 33,558
Todd M. Oberdick, Registration No. 44,268
Attorneys for Applicants
ST.ONGE STEWARD JOHNSTON & REENS LLC
986 Bedford Street
Stamford, CT 06905-5619
203 324-6155

09/09/05
Date Considered

Examiner

Notice of References Cited	Application/Control No. 10/822,310	Applicant(s)/Patent Under Reexamination LAVOIE ET AL.	
	Examiner Rachna Singh	Art Unit 2176	Page 1 of 2

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
*	A	US-6,792,145 B2	09-2004	Gay, Robert W.	382/190
*	B	US-2004/0230892 A1	11-2004	Horton, D. Troy	715/511
*	C	US-6,029,175	02-2000	Chow et al.	707/104.1
*	D	US-6,336,094 B1	01-2002	Ferguson et al.	705/1
*	E	US-6,324,555 B1	11-2001	Sites, Richard Lee	715/517
*	F	US-6,446,072 B1	09-2002	Schulze et al.	707/10
*	G	US-2004/0054967 A1	03-2004	Brandenberger, Sarah M.	715/511
*	H	US-2002/0116417 A1	08-2002	Weinberg et al.	707/517
*	I	US-5,956,726	09-1999	Aoyama et al.	707/102
*	J	US-2003/0084424 A1	05-2003	Reddy et al.	717/105
*	K	US-2002/0161860 A1	10-2002	Godlin et al.	709/219
*	L	US-2004/0220980 A1	11-2004	Forster, Karl J.	707/204
*	M	US-2005/0154664 A1	07-2005	Guy et al.	705/035

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Cooper, James, et al, Detecting Similar Documents Using Salient Terms, November -9, 2002 CIKM '02. Pages 245-251.
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Notice of References Cited	Application/Control No. 10/822,310	Applicant(s)/Patent Under Reexamination LAVOIE ET AL.	
	Examiner Rachna Singh	Art Unit 2176	Page 2 of 2

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
*	A	US-2006/0167772 A1	07-2006	Zilberman, Ran	705/035
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Detecting Similar Documents Using Salient Terms

James W. Cooper, Anni R. Coden, Eric W. Brown

IBM T J Watson Research Center
PO Box 704, Yorktown Heights, NY 10598
{jwcnmr, anni, brown} @ watson.ibm.com

ABSTRACT

We describe a system for rapidly determining document similarity among a set of documents obtained from an information retrieval (IR) system. We obtain a ranked list of the most important terms in each document using a rapid phrase recognizer system. We store these in a database and compute document similarity using a simple database query. If the number of terms found to not be contained in both documents is less than some predetermined threshold compared to the total number of terms in the document, these documents are determined to be very similar. We compare this to the shingles approach.

Categories and Subject Descriptors

H3.3 **[Information Search and Retrieval]** Clustering, Information Filtering, Selection process.

General Terms

Algorithms, Measurement, Documentation, Performance, Design

Keywords

Text mining, Duplicate documents, Databases, Shingles, Document similarity.

1. INTRODUCTION

One of the continuing problems in information retrieval is the fact that in the web environment, there are a large number of near-duplicate documents returned from most searches. A number of methods have been proposed for recognizing and eliminating such duplicates.

For example, Brown and Prager [1] note that documents having similar rank and identical metadata such as size, date, and base filename are likely to be copies kept on different directories or on different servers and can effectively be reduced to one single occurrence.

CIKM'02, November 4-9, 2002, McLean, Virginia, USA.

A more sophisticated system was described by Broder [2], in which regions of each document, called "shingles" are each treated as a sequence of tokens and then reduced to a numerical representation. These are then converted to "fingerprints" using a method originally described by Rabin [3]. Since the number of identical tokens could be compared, a document similarity measure could also be computed.

At a more simplistic level, Bloomfield [4] has recently described an algorithm for detecting plagiarism in which he simply searches for matches of six or more successive words between two documents. This provides as measure of identical embedded document sections, but not necessarily of document equality.

In this paper, we analyze the technique of characterizing documents using the major multi-word terms discovered in them using phrase recognition software, and develop the hypothesis that documents which have identical lists of discovered terms are effectively identical in content. In this context we use "terms" to mean multi-word terms and salient single word terms.

In the first section we discuss the text mining technology we use, and how we manage the data. In the next section we discuss how we define duplicate documents and how we detect them using the results of text mining. We describe the shingles algorithm and throughout compare this term-based technique with the shingles technique. Then in the following section, we propose a novel "document signature" for the rapid comparison of documents, and finally we discuss how this technique can be used quite successfully for finding documents that are variations on the original document.

2. THE TALENT TEXT MINING SYSTEM

In approaching document comparisons, we utilized the suite of text analysis tools collectively known as Talent (Text Analysis and Language Engineering Tools) for analyzing all the documents in the collection.

2.1 Textract Phrase Extraction

The primary tool we use for analyzing documents is **Textract**, itself a chain of tools for recognizing multi-word terms and proper names. We have described the features of Textract previously [5], [6]. Textract recognizes named entities [7], multi-word terms [8] and named [9] and unnamed relations between terms.

Textract reduces related forms of a term to a single *canonical form* that it can then use in computing term occurrence statistics

more accurately. In addition, it recognizes abbreviations and finds the canonical forms of the words they stand for and aggregates these terms into a vocabulary for the entire collection, and for each document, keeping both document and collection-level statistics on these terms. Each term is given a collection-level importance ranking called the IQ or Information Quotient [5], [14]. IQ is effectively a measure of the document selectivity of a particular term: a term that appears in only a few documents is highly selective and has a high IQ. On the other hand, a term that appears in many documents is far less selective and has a low IQ. We note that IQ may not be an especially significant metric across the entire web, where many domains are discussed, but within a single domain, or within the results of a specific search, it can be quite useful.

Textract further categorizes the entities it discovers into one of the categories shown in Table 1. The earlier categories have the least certainty and the later ones higher certainty.

Table 1 –Categories assigned to terms by Textract

UWORD	Unknown word
UTERM	Unknown term
UABBR	Unknown abbreviation
UNAME	Unknown type of name
PLACE?	Probably a place
PERSON?	Probably a person
PLACE	A place
PERSON	A person
ORG	An organization

While Textract is our tool of choice in these experiments, we note that there are a number of other systems that have been developed for phrase recognition. For example, Liddy has developed DR-LINK, which is marketed by Textwise [10], and Evans (et al) have developed LinkIT [11]. Further, somewhat similar technology is available in the ThingFinder program from InXight [12].

2.2 Data Management

Once the data from a collection of documents have been analyzed using Textract, it is useful to construct a system for managing that data in a way that makes it easy to sort the terms by document, by IQ and by term type. For this purpose, we constructed a Java class library known as KSS [15], [16] (for Knowledge System Server), which builds a database from the Textract data.

3. DUPLICATE OR SIMILAR DOCUMENTS

In any web search, it is fairly likely that some documents that are returned are very similar. For example, the same documents may exist on several servers or in several users' directories. Some very frequently found examples of this are the Java API documents from Sun, which are found on almost every Java developer's machine. Since these are very well known and described, they are very easy to eliminate using any of the existing techniques.

However, more difficult cases occur when

- There are several edited versions of the same document on various servers.
- The same document is found in several forms, such as HTML and PDF.
- A document is embedded in another. In this case the embedded document may or may not be the most significant part of the combined document.

These cases are more difficult to solve rapidly and it is these that make up the subject of this study.

For the purposes of this study, we define duplicate documents as ones that have essentially the same words in the same sentences and paragraphs. These paragraphs and their sentences could be in a somewhat different order, however. The documents may be in different forms, such as HTML and PDF and they may vary slightly in header boiler-plate information. We define *similar* documents as those in which a large percentage of the sentences, or words in the sentences, are the same.

4. THE SHINGLES TECHNIQUE

We will be contrasting our findings with those obtained using the well-known shingles technique [2]. This technique amounts to reducing each document to a series of numeric codes, such as hash codes, based on sequences of words. In the original paper, the authors suggested making each hash code of a group of 10 adjacent words, and moving the window by one word to create the next hash code. They then eliminate duplicates and, to reduce the number of values, save only those divisible by 25. If this is still too many, they save only the 400 smallest values. The advantage of using shingles to compare documents is that a simple set membership between two tables of integers can be computed very rapidly. Documents that match in all shingles are assumed to be identical and those that match nearly all shingles are closely related.

5. CURRENT EXPERIMENTS

Given the array of sophisticated term management technologies our group has developed, we undertook to find out whether these systems can be use to detect document similarity. In particular, is it possible to use some subset of terms found by Textract as a compressed document representation, which we can use to make rapid comparisons of documents and cluster or eliminate those that are essentially identical?

5.1 Query 1: Detecting Similar Documents

In our first experiment, we went to a popular search engine site with all enhancements turned off, and issued the query "fix

broken Thinkpad." This is the sort of naïve query that returns a plethora of documents the user does not want, or expect. Much as we predicted, there were *no* documents on how to repair Thinkpads. However, many of the top 50 documents contained all of these terms in some context or other. Of the top 50, we were able to download and analyze 36 of them. The remainder were broken links. Ten of these documents were Adobe Acrobat PDF files. We used the Gemini plug-in [17] for Adobe Acrobat Reader to export these files into HTML.

We then created a single collection of these documents and analyzed it using Textract. Textract produces a file of terms found in each document, and a file of all the terms in the collection along with their IQ. We used the KSS Java libraries to load these result files into DB2 and subjected the results to various SQL queries to determine the number of terms that documents had in common.

5.2 Similarity Queries

We first must find the significant terms in each document. Initially, we ranked all the terms except the unknown word types in order of decreasing IQ. We note that the IQ measure is a collection level statistic, and that this processing must be done on an entire group of returned documents for this measure to be meaningful.

The question we then want to ask, then, is which terms are not found in common between pairs of documents. You can find these in a single SQL query of the sort

```
Select count(*) as c from
  (select terms from doc1 where ..)
not in
    (select terms from doc2 where ..)
```

After some experimentation, we determined that the important selection criteria are to select terms with an IQ > 0 and which were not UWORDS.

This returns the count of the number of terms that appear in document 2 that are not in document 1. While it might seem that n^2 queries are necessary, it is really only necessary to traverse the upper triangle of this matrix. We do this by sorting the documents into order of increasing size, and comparing documents with the next larger ones in the list. The comparisons are done rapidly in a single database query, and we further reduce the number of compares by limiting the test to documents that are no more than about 10% larger than the compared document. This parameter is, of course, adjustable.

5.3 Results

We found 6 clusters of documents in the 36 documents we analyzed in the first query. Three of these were pairs of identical documents returned from different servers, as shown in Table 2. These documents are identical by any measure and can be easily recognized and collapsed to a single entity, using the methods described by Brown and Prager [1].

Table 3 contains an interesting cluster of eight documents that have similar names, but different sizes. The final two columns of the table shows the fractional difference in contained terms and shingles between adjacent documents in the list.

It is easy to see that these documents must be closely related versions of the same information. In fact, they are all different versions of the same IBM manual describing the Websphere server product. They differ in small details: for example one manual mentions the Linux version of Websphere and another does not. Each of these documents was returned as a PDF file and was converted to HTML using the Gemini plug-in mentioned above.

In Table 3, documents 1 and 2 and documents 3 and 4 are almost certainly absolutely identical. However, the remaining four documents are clearly all closely similar as well. This algorithm finds such documents much as shingles does, but with the added advantage that you can find out which salient terms appear in one and not in the other.

With this term information, one could thus remove some document from the cluster of very similar documents if the terms that are different between them include a term that is also contained in the original query.

Table 2– Identical documents returned by Query 1

Document name	Size	Term diff	Shingle diff
Sytpf130, Sytpfl1301	12236	0	0
aixy2k. aixy2k1	153255	0	0
Conf, client	52854	0	0

Table 3– Very Similar Documents returned from Query 1

#	Title	Size	Terms	Delta terms	Shingle Delta
1	Fund2	481963	2198	--	--
2	Fund4	481963	2207	0	0
3	ct7mhna1	493225	2139	0.014	0.112
4	ct7mhna2	493295	2146	0	0
5	Fund0	503138	2235	.029	0.074
6	Fund1	504737	2249	.016	0.120
7	Fund3	505004	2287	.011	0.094
8	Fund5	505471	2271	.005	0.012

This search also returned two other closely related document pairs as shown in Table 4. the two documents are in fact a draft and a

final PDF document of a paper by Selker and Burleson. [18] Since these papers are quite different in size and format, they would probably not have been found as similar by simple size and filename methods. The term differences between the two are partly because of some additional abstract and polishing, and partly because of the included boilerplate from the magazine format. However, it is surprising to note that the shingles technique does not suggest that these are similar at all, if we assume that a difference of 10% or less is required to call documents similar. We note that the papers each contain about 7800 single words after markup is removed, and that Textract finds 169 multi-words and about 1300 total terms in each of them.

Table 4 – Similar Documents from Query 1

Title	Size	Terms	Term diff	Shingles diff
Selhtm.htm	50169	257	--	--
Selpdf.htm	54274	218	.106	0.44

Careful examination of these two documents and their shingles led to three explanations for the differences:

1. The PDF document contains a table of contents and running page numbers which do not appear in the HTML version.
2. The hyphenation of the PDF document is different and more tokens are created containing trailing hyphens. We changed our shingling program to remove these hyphens and recombine the words, with little improvement.
3. The Gemini PDF translator converts hyphens to "soft hyphens" instead of "hard hyphens," and these have different character codes.

Replacing the hyphen character codes (0xad) with the hard hyphens (0x2d) resulted only in a slight change in the shingles similarity, reducing 0.44 to 0.39. Thus, the documents are genuinely different at the word-by-word level tracked by shingles, but the same using the salient term comparison approach.

6. DETECTING IDENTICAL DOCUMENTS

When documents are very large, it is not usually convenient to run phrase recognition software on the entire set of documents in real time when they are returned from a query, because the elapsed time is too great. However, as part of the indexing process, it is not unreasonable to catalog the major terms in each document. However, even making comparisons among large numbers of terms in multiple documents can take many seconds and can lead to unacceptable delays in returning answers.

We suggest that it is possible to compute a digital signature of each document, based on the terms we find in it. Such a signature can be as simple as a sum of the hash codes of the term strings that make up these major terms. In this experiment, we used the Java String class's hashCode method to compute a code for each term found in a document, and then added these codes up to form the signature. The results are shown in Table 5. We note that this simplification ignores the term order, and could in the pathological extreme possibly result in documents that differed by a negative stop-word such as "not," thus conceivably equating documents which reach opposite conclusions.

Table 5- Computed document signatures

Url	Size	Term Signature	Shingle signature
Fund2	481963	24681058826	**-846170517750**
Fund4	481963	26371644438	**-846170517750**
Ct7mhna1	493225	33130506660	***-852736297100***
Ct7mhna2	493295	32920545738	***-852736297100***
Fund0	503138	10451017932	-861022256825
Fund1	504737	8933271588	**-865769789850**
Fund3	505004	7211814280	**-865769789850**
Fund5	505471	12114748848	-861022256825
Sytpf130	12236	**13802917585**	**1506364450**
Sytpf1301	12236	**13802917585**	**1506364450**
aixy2k	153255	***28232730155***	***-128905766325***
aixy2k1	153255	***28232730155***	***-128905766325***
Client	52854	**6580188642**	**30847104000**
Conf	52854	**6580188642**	**30847104000**

This is in many ways the logical extreme of the "super-shingle" approach proposed by Broder *et. al.* [2], where the original shingle values are reshingled in groups of ten. The resulting much smaller table can be compared more quickly, but of course does not provide a way to detect similar documents. To provide an analogous number, we computed the shingles signature, by just adding up the shingles, shown in the last column of Table 5.

We note that while individual strings will usually have unique hash codes, there is a somewhat larger probability that the sum of a series of hash codes will be less unique. However, the probability of these collisions is small enough that these document signatures remain quite useful. Further, it is even less likely that documents with accidentally identical signatures would be returned from a query if they were not the same document.

To summarize, these document signatures simply provide a shorthand method of representing the top terms in documents, so

that they can be compared very rapidly. The actual technique for comparison is essentially the same as in section 5.1.

7. QUERY 2 – SMALLER DOCUMENTS

In a second series of experiments, we issued a more focused query "Program ViaVoice in Java," and were able to retrieve 47 of the top 50 returned documents. Since many of these had the same filename, we carefully renamed them when we saved the local copies for analysis.

Since all of these documents were of modest size (the top one was 75 K) we found that we could perform the entire analysis on the documents fast enough that it could be carried out more or less in real time in response to a query.

The term signature results included 8 pairs of identical documents as measured by size and the signature we described above. In addition, the results contained the 13 very similar documents shown in Table 6.

Table 6 –Very similar documents returned by Query 2

Url	Size	Textract Signature	Shingle signature
News11	37788	-9902679640	*-2788203200*
News9	38100	**-9692972733**	**-10719056625**
News5	38383	-11688727862	*-2788203200*
News12	38604	**-9692972733**	**-10719056625**
News8	38727	-9921821918	-1343956500
News2	39389	**-9692972733**	**-2654568325**
News6	39389	**-9692972733**	**-2654568325**
News3	39465	**-9692972733**	***-4657627900***
News4	39465	**-9692972733**	***-4657627900***
News0	39580	-5116033555	*-3060800550*
News1	39580	-8166342237	*-3060800550*
News10	40537	-11188846377	**-5207686200**
News7	40537	-12715476873	**-5207686200**

The documents in Table 6 are all very similar, since they differ in only 1 or 2 terms out of 47, and all have similar sizes. Based on size alone, you would identify only 4 pairs of identical documents. However, all of these are detected as similar based on the fact that they contain the same terms. In addition, it is significant that six of these documents have identical signatures (shown in boldface) even though they are of four different sizes. This shows the power of the signature method for rapid identification of documents.

By contrast, the shingles signatures identified 6 pairs of identical documents, but did not recognize that 4 of these pairs were effectively identical as the text signatures did.

8. FINDING SIMILAR DOCUMENTS

In the foregoing, we have discussed the problem of finding very similar documents, in most cases so similar that only one of them need be returned from a search at all.

There is another set of problems to solve, however, relating to documents that are similar because they exist in a number of revisions. In order to test this algorithm for this problem, we determined that we should relax the restrictions regarding the percentage of terms that could be different, and the size differences we would allow between documents we compare.

In the first experiment, we collected 12 documents about IBM financial and banking products and services from the ibm.com web site, so that they would all have a relatively similar vocabulary, and one document which was a press release about IBM's performance in a supercomputer comparison. We expected that this last document would have a markedly different vocabulary from the others.

Then we took this last supercomputer document and cut out a 2533 byte segment comprising the main news story without any of the surrounding HTML formatting, and pasted it into each of the other financial documents. We then ran Textract and indexed the terms per document as described above and ran the same experiment on document similarity, where we changed the SQL query to allow the fraction of different terms to be as high as 0.5. This query identified every document pair correctly and did not find any pairs of documents similar except those consisting of an original and the same document with inserted text.

Table 7 shows the fraction of terms that differ between the original document and the same document when the news release text is inserted. The fractional text differences vary between 0.01 and 0.157. However, here shingles did not do nearly as well, with only 5 of the 12 document pairs being identified as similar.

We concluded that documents that had less than 20% of the terms different were likely to represent documents that were related and contained much the same text. In fact, we found it quite encouraging that this method identified every such document correctly and returned no false positives. (A false positive would be a difference in terms of less than 20% in documents that were in fact different.) In other words the precision and recall for the text method were 100%, while for the shingles method, the recall was less than 50%.

Table 7 – Fractional differences in terms in financial documents when an news release on supercomputers was added to each of them

#	Original url	Fraction of different terms with inserted news release	Fraction different using shingles
21	Folder1	0.100	0.28
38	Reuters	0.157	0.41
30	Nletter	0.06	0.06

20	Folder	0.117	0.86
17	Ebus3	0.076	0.84
16	Ebus1	0.055	1.00
35	Retaildel	0.096	0.01
27	Kookmin	0.054	0.03
1	24552	0.040	0.48
8	Building	0.040	0.03
14	Ebusmark	0.010	.26
33	RetailB	0.015	0

On the other hand, neither algorithm identified the short IBM press release document as being related to any of the others by containment, since it was relatively short, and contained fewer salient terms.

9. IMPLEMENTATION DETAILS

In these experiments, we ran the Textract text mining program on the collection of documents (around 50) returned from the query. Then we generated low-level DB2 table load files [19] from the Textract output and loaded the terms/document data into DB2. The IQ and frequency of the terms was determined from this collection. Thus, IQ would change somewhat based on the contents of the documents returned. A term that was highly salient in one document set might appear too frequently to be very selective in another set. However, we have eliminated much of this dependence by simply requiring that the IQ value be non-zero. It would in general be possible to maintain a vocabulary for a search system with IQs predetermined.

When all of the documents are relatively short, it is quite possible to do this more or less in real time. However, when longer documents make the mining processes too slow, it is necessary to index and mine the documents in advance and cache the results, just as you do with the document search indexes. When database comparisons of strings in very long documents can be slow, it is possible to just compare the top terms, for example the top 200 terms in each document. Further, we can store a numeric hash code for each term which can be compared more rapidly.

Finally, it is quite reasonable to store the document signature we describe as part of the database document table, so you can compare documents quickly.

In the course of these experiments, we varied the IQ threshold and the term frequency threshold. For various types of applications, these values may well need to be adjusted. However, it is important to note that the document signature is dependent on the number of terms you retrieve, and if you change your criteria, you will need to recompute these signatures.

In comparing documents for close similarity as we did in Query 1, we only considered documents that were within 10% of the size of the one we were comparing to, and only considered documents to be similar when the number of terms that were different was less than 10% of the total number of terms in the smaller document. In comparing documents that contained embedded additional material, we relaxed both of these criteria to 50%, with little performance penalty.

9.1 The Shingles Implementation

For PDF documents, we used the Acrobat Gemini plugin to convert them to HTML. For each HTML document, we extracted the pure text using the Java HTML classes, or when that was unsuccessful, saving the files as text from Internet Explorer. We then converted the entire document to lower case. The shingles were calculated using a Java program by using a moving 10-word window, which computes the Java String hashcode for each 10 word token. The set of hash codes were exported into a DB2 load file, loaded into a DB2 table, and the unique values extracted into a second table. For longer documents, we limited the total number of shingles s generated to those where

$$s \; Mod \; 25 = 0$$

as the authors proposed. For very short documents we also repeated the comparisons using all the shingles, with no significant change in the results.

10. SUMMARY AND CONCLUSIONS

We define similar documents as ones that have essentially the same sentences and paragraphs, but not necessarily in exactly the same order. We have found that we can accurately compute whether documents are similar by comparing the number of terms found using a phrase recognition program such as Textract. Using this technique, we can also give the user a list of the terms by which similar documents differ.

We further found that you can accurately recognize documents that have been revised to contain parts of other documents as still being closely related to the parent document. Finally, we described a novel document signature that you can use to make a rapid comparison between documents that are likely to be identical.

We contrast the success of this method with the shingles method as follows. For documents from the same source and file format, the techniques largely give the same results. However, when documents that originate in different file formats (such as PDF and HTML) are compared, the term-based method appears to be more successful. Further, while it would be possible to do some post-processing on translated PDF files before shingling them to try to eliminate differences, this amounts to the same sort of linguistic processing our technique already employs.

In the case of comparing documents with inserted text, the term method seems to be more successful, since its recall was much higher.

We further note that our term method uses only the salient terms to characterize documents, and these terms can appear in a different order and still provide the same characterization of the

document. Further, phrase recognition programs such as Textract generally reduce the found terms to a root or "canonical" form, so that even if the terms appear in different variant forms in slightly edited versions of a document, they will be recognized as being the same root term and found to be identical. Finally, this method is insensitive to the addition of additional polishing sentences or the rearrangement of whole paragraphs in edited versions of a document.

This system has broad applicability in improving the results of searches of large document collections, whether the returned documents have been indexes for their term content in advance or not. It can also be used for rather sophisticated plagiarism detection, or as an adjunct in finding further documents of interest and grouping these documents for the user's convenience.

11. ACKNOWLEDGEMENTS

We thank Geoffrey Zweig of IBM Research for helpful discussions, and Robert Mack, Roy Byrd and Alan Marwick for their support.

12. REFERENCES

[1] Brown, Eric W. and Prager, John M., US Patent 05913208.

[2] Broder, Andrei Z, Glassman, Steven C., Manasse, Mark and Zweig, Geoffrey "Syntactic Clustering of the Web," *Proceedings of the Sixth WWW Conference.* Santa Clara, CA, 1997.

[3] Rabin, M. O., "Fingerprinting by random polynomials, " Center for Research in Computing Technology, Harvard University, Report TR-15-81, 1981.

[4] Bloomfield, Louis, University of Virginia, interviewed on NPR's *All Things Considered*, May 9, 2001. See www.plagiarism.phys.virginia.edu.

[5] Cooper, J. W. and Byrd, R J, "Lexical Navigation: Visually Prompted Query Refinement," ACM Digital Libraries Conference, Philadelphia, 1997.

[6] Cooper, James W. and Byrd, Roy J., OBIWAN – "A Visual Interface for Prompted Query Refinement," Proceedings of HICSS-31, Kona, Hawaii, 1998.

[7] Ravin, Y. and Wacholder, N. 1996, "Extracting Names from Natural-Language Text," IBM Research Report 20338.

[8] Justeson, J. S. and S. Katz "Technical terminology: some linguistic properties and an algorithm for identification in text." *Natural Language Engineering,* 1, 9-27, 1995.

[9] Byrd, R.J. and Ravin, Y. Identifying and Extracting Relations in Text. *Proceedings of NLDB 99*, Klagenfurt, Austria.

[10] Mnis-Textwise Labs, www.textwise.com. DR-LINK was developed at Syracuse University and is marketed by Textwise.

[11] Evans, D. K., Klavans, J. and Wacholder, N., "Document Processing with LinkIT," *Proc. Of the RIAO Conference,* Paris, France, 2000.

[12] InXight, Inc. www.inxight.com

[13] Neff, Mary S. and Cooper, James W. "Document Summarization for Active Markup," *in Proceedings of the 32nd Hawaii International Conference on System Sciences,* Wailea, HI, January, 1999.

[14] Cooper J.W. and Prager, John M. "Anti-Serendipity – Finding Useless Documents and Similar Documents," *Proceedings of the 33rd Hawaii International Conference on System Sciences, Maui, HI,* January, 2000.

[15] Cooper, J. W. "The Technology of Lexical Navigation," Workshop on Browsing Technology, *First Joint Conference on Digital Libraries,* Roanoke, VA, 2001.

[16] Cooper, J.W., Cesar, C., So, Edward, and Mack R. L., "Construction of an OO Framework for Text Mining," *OOPSLA,* Tampa Bay, 2001.

[17] Gemini plug-in for Adobe Acrobat Reader, Iceni Technology, Ltd, Norwich, England, www.iceni.com.

[18] Selker, T. and Burleson, W. "Context-aware Design and Interaction in Computer Systems," *IBM Systems Journal,* **39,** 891 (2000).

[19] Cooper, J W, "Loading Your Databases," *JavaPro,* May, 2000.

Organization TC2100 Bldg./Room RANDOLPH
U. S. DEPARTMENT OF COMMERCE
COMMISSIONER FOR PATENTS
P.O. BOX 1450
ALEXANDRIA, VA 22313-1450
IF UNDELIVERABLE RETURN IN TEN DAYS

OFFICIAL BUSINESS

UNITED STATES POSTAGE
02 1A
0004204479
MAILED FROM ZIP

AN EQUAL OPPORTUNITY EMPLOYER

LAV0396 021163360 1905 11 10/02/06
RETURN TO SENDER

LAVOIE
MOVED LEFT NO ADDRESS
UNABLE TO FORWARD
RETURN TO SENDER

RECEIVED
OCT 11 2006
USPTO MAIL CENTER